



## UNITED STATES AIR FORCE RESEARCH LABORATORY

---

### SALVAGING CONSTRUCT EQUIVALENCE THROUGH EQUATING

Malcom James Ree

CENTER FOR LEADERSHIP STUDIES  
OUR LADY OF THE LAKE UNIVERSITY  
411 SW 24th STREET  
SAN ANTONIO TX 78207

Thomas Carretta

HUMAN EFFECTIVENESS DIRECTORATE  
CREW SYSTEM INTERFACE DIVISION  
WRIGHT-PATTERSON AFB OH 45433-7511

James A. Earles

AIR FORCE OCCUPATIONAL MEASUREMENT  
SQUADRON  
RANDOLPH AFB TX 78150

20000112 080

AUGUST 1999

INTERIM REPORT FOR THE PERIOD JULY 1998 TO DECEMBER 1998

Approved for public release; distribution is unlimited

Human Effectiveness Directorate  
Crew System Interface Division  
2255 H Street  
Wright-Patterson AFB OH 45433-7022

## NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service  
5285 Port Royal Road  
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center  
8725 John J. Kingman Road, Suite 0944  
Ft. Belvoir, Virginia 22060-6218

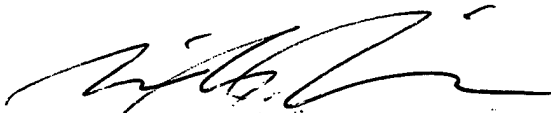
## TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-1999-0187

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

**FOR THE COMMANDER**



MARIS M. VIKMANIS  
Chief, Crew System Interface Division  
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE August 1999	3. REPORT TYPE AND DATES COVERED Interim - July 1998 to December 1998	
4. TITLE AND SUBTITLE  Salvaging Construct Equivalence Through Equating			5. FUNDING NUMBERS  PE - 62202F PR - 1123 TA - B1 WU - 01	
6. AUTHOR(S) Malcolm James Ree* Thomas R. Carretta** James A. Earles***				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Center for Leadership Studies*      Air Force Occupational Measurement Our Lady of the Lake University      Squadron*** 411 SW 24th Street      RandolphAFB, TX 78150 San Antonio, TX 78207			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory** Human Effectiveness Directorate Crew Systems Interface Division Air Force Materiel Command Wright-Patterson AFB, OH 45433-7511			10. SPONSORING/MONITORING AGENCY REPORT NUMBER  AFRL-HE--WP-TR-1999-0187	
11. SUPPLEMENTARY NOTES  Air Force Research Laboratory Technical Monitor: Dr. Thomas R. Carretta, (937) 656-7014; DSN 986-7014				
12a. DISTRIBUTION AVAILABILITY STATEMENT  Approved for public release; distribution is unlimited			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words)  Turban, Sanders, Frances, and Osburn (1989) provided a two-step procedure for selecting a replacement for a current personnel test without an expensive validation study. The two steps are confirmatory factor analysis and impact analysis. It is possible to apply the two-step procedure and find that the replacement test is not acceptable. We provide an example of just such a negative outcome that was salvaged by the extra step of equipercentile equating. This step, added to Turban et al., required no additional investment other than an equating analysis on the extant data.				
14. SUBJECT TERMS Air Force Officer Qualifying Test      Ability measurement Scholastic Achievement Test Test equating			15. NUMBER OF PAGES 17	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT  Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE  Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT  Unclassified	20. LIMITATION OF ABSTRACT  UL	

THIS PAGE INTENTIONALLY LEFT BLANK

## **PREFACE**

This effort was performed under work unit 1123-B1-01 in support of USAF aircrew selection and classification. The authors thank A. Berndt, D. deBoom, N. Dorans, H. Read, P. Rioux, and E. Wolfe for their help in this effort. Send e-mail to the first author at [MREE@STIC.NET](mailto:MREE@STIC.NET).

## TABLE OF CONTENTS

	Page
SUMMARY .....	1
INTRODUCTION .....	1
METHOD .....	3
Participants .....	3
Measures .....	3
Analyses .....	3
RESULTS AND DISCUSSION .....	7
REFERENCES .....	11

## FIGURES

	Page
1 Structural Models of the Composites .....	6

## TABLES

	Page
1 Descriptive Statistics and Correlations for the Composite Scores .....	8
2 Confirmatory Factor Analyses of the Relation between AFOQT and SAT .....	9
3 Chi-Square Impact Analyses of the Two-Test Verbal-Quantitative Composite Equated by Linear and Equipercentile Methods .....	10

# SALVAGING CONSTRUCT EQUIVALENCE THROUGH EQUATING

## SUMMARY

Turban, Sanders, Frances, and Osburn (1989) provided a two-step procedure for selecting a replacement for a current personnel test without an expensive validation study. The two steps are confirmatory factor analysis and impact analysis. It is possible to apply the two-step procedure and find that the replacement test is not acceptable. We provide an example of just such a negative outcome that was salvaged by the extra step of equipercentile equating. This step, added to Turban et al., required no additional investment other than an equating analysis on the extant data.

## INTRODUCTION

Sometimes it is necessary to discontinue the use of a validated test and substitute another. The need for a replacement test could arise from compromise, obsolescence, cost constraints, or source unavailability. Obviously the replacement test must be valid, but validation studies are very expensive (Seberhagen, 1996). Turban, Sanders, Frances, and Osburn (1989) suggested a two-step procedure for picking a replacement test without an expensive validation study. First, confirmatory factor analytic methods are used to determine the level of construct equivalence. Turban et al. proposed testing three models of construct equivalence by constraining and releasing parameters estimated in confirmatory factor analyses. The first model investigates congeneric equivalence, which is whether the tests measure the same construct. The second more restrictive model is *Tau* equivalence and means that the tests measure the same construct and the distribution of true scores is equal for both tests. The third and most restrictive model is parallel equivalence where the tests measure the same constructs, have equal true score distributions, and have equal error score distributions. Two parallel equivalent tests can be used interchangeably and will have equal validity.

The second step is impact analysis. This is an investigation of cutoff scores to determine the consequences of substituting one test for another. For example, does an existing cut-off score, transferred to a new test, qualify the same proportion of applicants as when used on the previous test? Impact analysis as proposed and conducted by Turban et al. (1989) computes a two-by-two table of passing both tests, failing both tests, passing one and failing the other, and failing one and passing the other. If impact analysis shows differing proportions qualified on the existing and replacement test, or if the passing/failing and failing/passing proportions are not the same, the two tests cannot be used interchangeably.

Anastasi (1989) suggested that Turban et al.'s (1989) innovative two-step approach be investigated and Hattrup, Schmitt, and Landis (1992) published an application. Although, Turban et al. and Hattrup et al. were successful in applying the method, this success is not necessarily guaranteed. Our attempt to apply the Turban et al. two-step approach led us to accept the factor analytic equivalence of two tests. Having found this equivalence in the confirmatory factor analyses, we expected further confirmation from impact analysis. We did not find it and were in the apparent position of having to abandon the effort and start over with a new candidate test.

As observed by Turban et al. (1989), when cutoff scores of construct equivalent tests are carried over from the old test to the replacement test, it is important that these carried-over cutoff scores retain their previous meaning. This will happen for all scores only if the distributional shapes of the tests are the same, in which case z-score transformation, as represented by the fully standardized output of a confirmatory factor analysis, would be appropriate. In a z-score transformation, also called a linear equating (Angoff, 1971; Flanagan, 1951), only the means and standard deviations of the distributions are set equal. When the shapes of the distributions are different the tests must be equated in some other way, the most common of which is the method of equipercentiles (Angoff, 1971).

The definition of equipercentile equating is that a score on test X is considered equated to a score on test Y when their percentile ranks are equal in a given group (Angoff, 1971; Flanagan, 1951). Equating can be performed either with one group of examinees or two groups depending on real-world constraints. Operationally this means that the tests are administered and percentiles are computed for each test. Scores on each test that have the same percentile rank are considered equated.

If only z-score transformations (standardized LISREL output) are used, as in Turban et al. (1989) and Hattrup et al. (1992), the confirmatory factor analytic results may lead to erroneous conclusions about interchangeability. First, it might be declared that a replacement test is not acceptable when equipercentile equating could make it acceptable. Second, as we demonstrate empirically, because of the fallible use of confirmatory factor analytic goodness-of-fit indexes, a replacement test might be declared interchangeable using the standardized output of LISREL, when it is not as shown by impact analyses.

In the present study, we investigated the practical problem of replacing an expensive proprietary test, the Air Force Officer Qualifying Test (AFOQT) with a commercially available test, the Scholastic Aptitude Test (SAT). The cost of developing and maintaining the AFOQT is great. It requires test booklets, answer sheets, answer keys, administration locations, and other substantial costs. Substituting at least the verbal and quantitative portions of the SAT for the verbal and quantitative portions of the AFOQT promised to yield substantial cost reductions. Although the SAT and the AFOQT were not written for the same purpose, the frequency of SAT administration, the common content, and a preliminary study (Ree & Carretta, 1998) that yielded correlations from .6 to .8 suggested the potential for substitutability.

We provide an example of the Turban et al. (1989) two-step procedure that required equipercentile equating. Had the step of equipercentile equating been overlooked, we would have declared the replacement test unusable and abandoned it. The added step of equipercentile equating salvaged the research investment in the replacement test.

## METHOD

### *Participants*

The participants were 7,940 men and women enrolled in the U. S. Air Force Reserve Officer Training Corps (ROTC). All were either in college or applying to college when they were



administered the proprietary test and applying to college when they were administered the commercial test. All test administrations took place between 1991 and 1994.

The average age of the participants was 19 years and 78 percent were male. The race/ethnic proportions were White, 82 percent, African-American, 7 percent, and other, 11 percent. Because the participants were subjected to prior selection, the scores created a range-restricted sample (Ree, Carretta, Earles, & Albert, 1994). Prior selection was on the basis of admission to ROTC including acceptance into college and scores on the proprietary test.

### *Measures*

#### *Air Force Officer Qualifying Test*

The proprietary AFOQT is a multiple-aptitude battery used for selecting officers. It is constructed, administered, and maintained by the U. S. Air Force. Its factor structure (Carretta & Ree, 1996) and its validity (Carretta & Ree, 1995) have been examined. Depending on specific occupational assignment, there are various minimum qualification scores on a verbal, quantitative, and/or a combined verbal-quantitative composite.

The verbal composite tests are Verbal Analogies, Reading Comprehension, and Word Knowledge. The quantitative composite tests are Arithmetic Reasoning, Data Interpretation, and Mathematics Knowledge. The combined verbal-quantitative composite uses all six tests. Raw scores are summed and converted to normative percentiles. The test-retest reliabilities of the AFOQT composites are: verbal .88, quantitative .84, and combined verbal/quantitative .88 (Carretta & Ree, 1997).

#### *Scholastic Aptitude Test*

The SAT, developed by the Educational Testing Service, Inc., provides verbal and quantitative scores used by many colleges and universities in their admissions process. The internal consistency reliabilities of the SAT are verbal .93 and quantitative .92 (Donolon & Livingston, 1984, pp. 33-34). We created a combined SAT verbal-quantitative composite through simple addition for use in later analyses.

### *Analyses*

When a new test replaces an old test and it is required to maintain validity, several steps must be followed. These steps are based in psychometric true-score theory (See Stanley, 1971, p. 369). First, it is necessary to compare the distributional shapes to determine if equipercentile equating is needed. The appropriate analyses must include an investigation of the score distributional shape parameters of skew and kurtosis (see Kendall & Staurt, 1977, p. 258 for standard errors) to determine if the standardized results from a confirmatory factor analysis can properly equate the tests. Angoff (1982) noted that standardization (generally called linear or z-score equating) converts only the mean and standard deviations of the distributions and assumes that the skew and kurtosis are the same for both tests. If they are not the same, equipercentile equating must be conducted. An equipercentile equating "...operates to match all moments (i.e., all characteristics

of the shape, in addition to the mean and standard deviation)...” (Angoff, 1982, p. 56). The standardized results of a confirmatory factor analysis will not necessarily provide any information about the skew and kurtosis of the tests. Interpretation of the standardized results might lead to the conclusion that two tests are interchangeable when they are not because of score distribution shape differences. The determination of the need for equipercentile equating is the first step and must be added to the Turban et al. (1989) procedure.

Second, the scores must be correlated. If the tests are not highly correlated, they cannot be interchangeable. Because the participants had been previously selected on these scores, they constituted a range-restricted sample relative to applicants and the correlations were downwardly biased. The correlations among the scores were corrected for range restriction using the multivariate method (Lawley, 1943; Ree et al., 1994; computer program available, see Johnson & Ree, 1994). A sample of 3,000 Air Force commissioning applicants (Skinner & Ree, 1987) provided the unrestricted values on the AFOQT for range-restriction correction. The multivariate method of correction for range restriction is the general solution to the problem and works from the variance-covariance matrix of all the variables (Ree et al., 1994). The more familiar equations called Cases 1, 2, and 3 (Ree et al., 1994) are specific simple forms of Lawley’s procedure. All make the same assumptions of equality of errors of estimate and constant regression form for the range restricted sample and the unrestricted sample. Lawley’s procedure is mathematically cumbersome, requiring matrix algebra and use of the Johnson and Ree program is advantageous.

Third is the accomplishment of an equipercentile equating (Angoff, 1971, 1982) when the skew and kurtosis differ. Equipercentile equating estimates the raw and true scores in two test score distributions that have the same percentile value if the reliability of the two test scores is about the same. If the reliabilities differ markedly a process called “regressed equipercentile equating” is conducted. This procedure estimates the true scores for the two distributions by Kelley’s equation (Kelley, 1947, p. 409) which is an estimate of true ability given a fallible measure. Kelley’s equation is:

$$X_T = r_{xx'}X_1 + (1 - r_{xx'}) M_1 \quad (\text{Eq. 1})$$

Where  $X_T$  is the true score for the individual,  $r_{xx'}$  is the reliability of the test,  $X_1$  is the individual’s test score, and  $M_1$  is the mean test score. Kelley observed:

This is an interesting equation in that it expresses the estimate of true ability as a weighted sum of two separate estimates, - one based upon the individual’s observed score,  $X_1$ , and the other based upon the mean of the group to which he belongs,  $M_1$ . If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice-versa.” (p. 409)

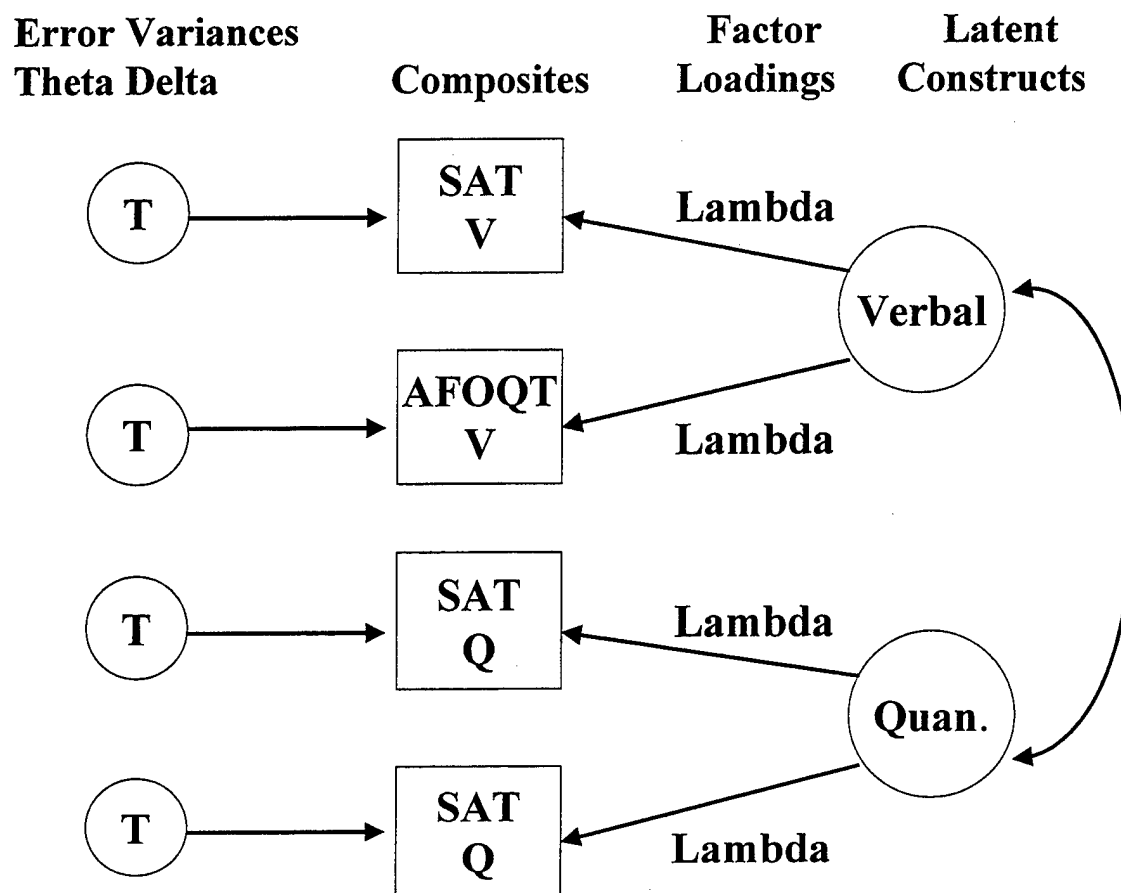
Replacing an operational test of high reliability with a new test of markedly lower reliability should not be considered without a validity study.

Equating is done by estimating local percentiles for both distributions and matching scores that have the same percentile equivalent. For example, there is a test score in each distribution that has a percentile equivalent of 1 and those two test scores are considered equivalent. There is

a test score in each distribution that has a percentile equivalent of 2 and those two scores are considered equivalent and so on up to the 99<sup>th</sup> percentile. The number of percentile points estimated is a function of the number of test questions. In a short test, a few percentiles (10 to 20) can be estimated and in a long test, many percentiles (50 to 99) can be estimated. Equipercentile equating produces an area transformation of the data. The shapes of the distributions of the original test and the equated test will be identical after equipercentile equating. Equipercentile equating provides a conversion table that matches scores on one test to scores on the other test. The equating function in this table is generally irregular because of sampling, but can be smoothed to produce a useable product. Smoothing of the equating table can be accomplished by function fitting, as we did in the current study, or by regression or other analytic means. The final result is a table that shows which scores are equivalent. Although this can be done by hand with graph paper and calculator, we conducted these equatings using proprietary software, avoiding computational inaccuracy and subjectivity in smoothing.

We replaced the SAT scores with the value of the AFOQT scores that equate to the SAT scores. This places the scores of the SAT composites on the same score scales as the composites of the AFOQT. That is, the equated SAT scores have the same metric, distributional statistics, and shape as the AFOQT scores.

Fourth, we followed the first step of the Turban et al. (1989) procedure for determining construct equivalence of the two verbal and two quantitative composites. This procedure tests three levels of equivalence based on the equality of latent constructs, true score variances, and error variances. Figure 1 presents the models of the three levels of equivalence. The least restrictive form, congeneric equivalence, requires that the tests measure the same latent constructs. That is, the lambdas must be non-zero. *Tau*-equivalence requires that congeneric tests measure the same latent constructs and have equal true score variances (Lambdas). Parallel equivalence adds the requirement of equal error variances (Theta-deltas) to the definition of *Tau*-equivalence. Further, not noted by Turban et al., parallelism requires that the test distributions have the same shapes (equal skewness and equal kurtosis). Tau -equivalence and parallel equivalence are successively more restrictive cases of congeneric equivalence. Neither congeneric nor Tau-equivalence requires similarity of distributional shapes (Lord & Novick, 1968, chapter 10).



**Figure 1. Structural models of the composites.**

Note. Null model: all factor loadings and phi equal zero.

Congeneric model: not the Null model. At least one pair of lambdas non-zero and have the same sign. No equality constraints on the lambdas nor the theta-deltas.

Tau equivalent model: by pairs, the lambdas are equal

Parallel model: by pairs the content equivalent lambdas are equal and the theta-deltas are equal.

Following Turban et al. (1989), the corrected-for-range-restriction correlations were disattenuated for unreliability. Using the formula given in Gulliksen (1950, p. 124, 1987, p. 124), we corrected the sample-dependent reliabilities to the values that would have been observed in the unrestricted sample. That is, the verbal-to-verbal, quantitative-to-quantitative, and verbal-to-quantitative correlations were corrected for range restriction and then disattenuated by reliabilities that were also appropriately corrected. These corrected correlations for verbal-to-verbal and quantitative-to-quantitative should approach 1 if the observed scores are measures of the same construct. Even if the corrected correlations were very high, we would not know the nature of the equivalence. Therefore, confirmatory factory analytic models for congeneric, Tau, and parallel equivalence were tested (LISREL 8.14, Jöreskog & Sörbom, 1996).

Confirmatory factor analyses using maximum likelihood estimation were performed on three matrices. The first analysis used the variance-covariance matrix of the observed scores and

retained the original (unequated) metrics and score distribution shapes of the composites. The second analysis used the standardized variance-covariance (correlation) matrix of the scores and is equivalent to applying z-score linear equating. The means and standard deviations are set equal but the shapes are not set equal. Finally, the third analysis, not conducted by Turban et al. (1989), used the variance-covariance matrix after the SAT was equated to the AFOQT by the method of equipercentiles.

We computed the chi-square, Comparative Fit Index (CFI), and the Root Mean Square Residual (RMSR) as our measures of model fit. This was similar to Turban et al. (1989) except that we substituted the smaller-sampling-variance CFI for their use of the Goodness of Fit Index. At the time of Turban et al. the CFI was not yet in wide use.

Finally, impact analyses as described by Turban et al. (1989) were conducted. A verbal-quantitative composite of the SAT scores and a verbal-quantitative composite of the AFOQT scores were created by addition. We specified three cutoff scores on the old composite, the AFOQT, to simulate high (80th percentile), moderate (50th percentile), and low (20th percentile) selectivity. At each of the three specified cutoff scores, counts were made of the number who qualified on both, one or the other, or neither verbal-quantitative composite, producing a two-by-two table. Random measurement error, unreliability, would cause some people to pass one composite but not the other composite. Chi-square statistics for goodness-of-fit were computed to determine if the numbers who passed one composite but failed the other differed. Type I error rate was set at  $p < .01$ . The chi-square analysis was done for the composites after linear (z-score) equating and again after equipercentile equating.

## RESULTS AND DISCUSSION

Table 1 provides the sample mean, standard deviation, skew, and kurtosis for each composite. The two SAT composites are mesokurtic and nearly normal, while the two AFOQT composites are platykurtic. There are also notable differences in skew. A statistical test may be conducted, (see McNemar, 1969, pp. 88-89), but given the large sample size most null hypotheses will be rejected. The composites differ sufficiently to make linear equating (z-score) inappropriate (Angoff, 1971, 1982). Therefore, for these composites, it is necessary to do an equipercentile equating regardless of the outcome of the standardized results of the LISREL analysis.

The observed and corrected-for-range-restriction correlations are also presented in Table 1. In range-restricted form, all the means were above applicant values and the standard deviations were reduced to about 80 percent of applicant values. The intercorrelations were all positive for both the uncorrected and the corrected-for-range-restriction matrices. As would be expected in a selection setting, the corrected correlations were larger than the uncorrected correlations.

**Table 1**  
**Descriptive Statistics and Correlations for the Composite Scores**

Descriptive Statistics

Composite	Mean	SD	Skew	Kurtosis
SAT-V	531.2	82.4	0.016	2.944
SAT-Q	605.5	89.3	-0.383	2.993
AFOQT-V	61.7	22.5	-0.200	2.047
AFOQT-Q	65.9	21.4	-0.539	2.405

Correlation Matrix

Composite	SAT-V	SAT-Q	AFOQT-V	AFOQT-Q
SAT-V	1.000	.617	.821	.571
SAT-Q	.538	1.000	.545	.823
AFOQT-V	.761	.441	1.000	.597
AFOQT-Q	.470	.752	.508	1.000

Note. SAT-V and SAT-Q are the verbal and quantitative composites of the Scholastic Aptitude Test. AFOQT-V and AFOQT-Q are the verbal and quantitative composites of the Air Force Officer Qualifying Test. Normal skew and kurtosis are 0 and 3.000, respectively. Entries below the diagonal of the correlation matrix are observed and those above have been corrected for range restriction.

The corrected-for-range-restriction correlations of the composites were then corrected (disattenuated) for unreliability. The fully corrected correlations between SAT and AFOQT were: .91 for the verbal composites and .94 for the quantitative composites, suggesting a near identity of the constructs measured.

The confirmatory factor analysis models of congeneric, *Tau*, and parallel equivalence were estimated on the observed, standardized, and equated score covariance matrices and are presented in Table 2. These analyses determine the construct equivalence of the composites.

The differences between the null model  $\chi^2$  and those for the specified models were quite large and all the  $\chi^2$  values were significant as would be expected with large samples. For the observed score AFOQT and observed score SAT, comparison of the CFI and RMSR goodness-of-fit indexes for the three models showed congeneric equivalence, but not *Tau* or parallel equivalence. Noteworthy is the poor fit of the parallel equivalence model (CFI = 0.0, RMSR = 3,201.32) as should be expected for tests that have different score scales and different shaped distributions (Angoff, 1971; Gulliksen, 1950, 1987).

**Table 2.**  
**Confirmatory Factor Analyses of the Relation Between AFOQT and SAT**

Model	$\chi^2$	df	p <	Comparative Fit Index	Root Mean Square Residual
Observed Score AFOQT And Observed Score SAT					
Null	22,772.854	6	.01	----	----
Parallel	36,901.191	5	.01	.000	3,201.032
Tau-equivalent	3,758.647	3	.01	.835	1,970.956
Congeneric	638.381	1	.01	.972	62.040
Standardized Score AFOQT And Standardized Score SAT					
Null	22,772.850	6	.01	----	----
Parallel	662.590	5	.01	.971	.017
Tau-equivalent	653.716	3	.01	.971	.018
Congeneric	638.381	1	.01	.972	.016
Observed Score AFOQT And Equated Score SAT					
Null	23,062.766	6	.01	----	----
Parallel	641.899	5	.01	.972	12.652
Tau-equivalent	630.586	3	.01	.973	13.767
Congeneric	613.496	1	.01	.973	11.377

Although previous analyses of skew and kurtosis have ruled out interpretation of the standardized results from the confirmatory factor analysis, we provide these results to make a point. The standardized analysis will not necessarily allow appropriate inferences about construct equivalence. As shown in Table 2, the CFI and RMSR for all three standardized models are nearly the same and indicate a good fit for congeneric, *Tau*, and parallel equivalence. This acceptance of parallel equivalence is in error because of the differing shapes of the distributions and would lead to unjustifiably replacing an old test with a new test. We were misled when following the Turban et al. (1989) two-step procedure in our initial analysis of construct equivalence because we did not evaluate skew and kurtosis and were dismayed when the follow-on impact analyses did not allow us to accept the replacement test.

Because of the differences in skew and kurtosis, an equipercentile equating was accomplished. After equating, we tested the three structural models again. Table 2 shows that there are no practical differences among the values of the fit statistics. After equipercentile equating, which equalizes the score scales and equalizes the distributional shapes, the composites are parallel equivalent. This is to be expected. Congeneric tests with similar reliabilities will be parallel if equated by the method of equipercentiles.

Impact analyses for high, moderate, and low levels of selectivity are presented in Table 3. These analyses were conducted for the verbal-quantitative composites as equated by linear and equipercentile methods. The  $\chi^2$  values for the linearly-equated composites were all statistically

significant. None of the  $\chi^2$  values for the equipercentile-equated composites was statistically significant. When the composites were equated with the equipercentile method, as many applicants passed the AFOQT and failed the equated SAT as passed the equated SAT and failed the AFOQT. The results of the impact analysis for the AFOQT and the equated SAT now allow us to accept the equated SAT as a replacement test. This was not true when the tests were equated by the linear method. Equipercentile equating salvaged the (equated) SAT as a valid replacement test.

**Table 3.**  
**Chi-Square Impact Analyses of the Two Test**  
**Verbal-Quantitative Composites Equated by**  
**Linear and Equipercentile Methods**

Percentile	Linear $\chi^2$	Equipercentile $\chi^2$
20	29.137**	.837
50	33.983**	.434
80	76.647**	.006

\*\*  $p < .01$

Why did confirmatory factor analysis show two composites to be parallel equivalent when they were not? It is because parallelism requires the distributional shape parameters, skew and kurtosis, to be equivalent for both composites. Confirmatory factor analytic methods do not use skew and kurtosis information. They all use covariance or correlation matrices and make assumptions about distributional shapes. Total reliance on the results of the confirmatory factor analysis is unwise. This is why we recommend a first step of distributional shape analysis and why Turban et al. (1989) recommend the step of impact analysis.

The equipercentile-equated SAT can be used to replace the verbal and quantitative AFOQT scores. They measure the same construct, are actually construct parallel, and perform equivalently in impact analyses. We have gone from failure to find interchangeability through the application of the unmodified Turban et al. (1989) two-step procedure to successful interchangeability through the salvaging step of equipercentile equating.



## REFERENCES

- Anastasi, A. (1989). Ability testing in the 1980's and beyond: Some major trends. *Public Personnel Management*, 18, 471-485.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). *Educational Measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council On Education.
- Angoff, W. H. (1982). Summary and derivation of equating used at ETS. In P. W. Holland & D. B. Rubin (Eds.). *Test equating*. (pp. 55-57). NY: Academic Press.
- Carretta, T. R., & Ree, M. J. (1995). Air Force Officer Qualifying Test validity for predicting pilot training performance. *Journal of Business and Psychology*, 9, 379-388.
- Carretta, T. R., & Ree, M. J. (1996). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology*, 8, 29-42.
- Carretta, T. R., & Ree, M. J. (1997). *The best retest score is the average: Findings and implications* (AL/HR-TP-1996-0021). Brooks AFB, TX: Manpower and Personnel Research Division, Armstrong Laboratory Human Resources Directorate.
- Donolon, T., & Livingston, S. (Eds.) (1984). Psychometric methods used in the admissions testing program. In T. Donolon (Ed.). *College board handbook for the Scholastic Aptitude Test and achievement tests* (pp. 33-34). NY: College Entrance Examination Board.
- Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.) *Educational measurement* (pp. 695-763). Washington, DC: American Council On Education.
- Gulliksen, H. (1950). *Theory of mental tests*. NY: Wiley.
- Gulliksen, H. (1987). *Theory of mental tests*. Mahwah, NJ: Erlbaum.
- Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially available aptitude tests. *Journal of Applied Psychology*, 77, 298-308.
- Johnson, J., & Ree, M. J. (1994). RANGEJ: A Pascal program to compute the multivariate correction for range restriction. *Educational and Psychological Measurement*, 54, 693-695.
- Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Chicago, IL: Scientific Software International.
- Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge MA: Harvard University Press.
- Kendall, M., & Staurt, A. (1977). *The advanced theory of statistics, Volume 1*,

(4<sup>th</sup> ed.), p. 258. NY: Macmillan.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh* (Section A, Part 1), 28-30.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McNemar, Q. (1969). *Psychological statistics* (4<sup>th</sup> ed.). NY: Wiley.

Ree, M. J., & Carretta, T. R. (1998). *Interchangeability of verbal and quantitative scores for personnel selection: An example*, AL/HR-TP-1997-0016. Brooks AFB, TX: Human Effectiveness Directorate, Air Force Research Laboratory.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology*, 79, 298-301.

Seberhagen, L. W. (1996). How much does a test validation study cost? In R. S. Barrett, (Ed.) *Fair employment strategies in human resource management*. Westport, CT: Quorum Books.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O*, AFHRL-TR-86-68. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Stanley, J. C. (1971). Reliability, In R. L. Thorndike (Ed.). *Educational Measurement* (2nd ed.) (pp. 356-442). Washington, DC: American Council On Education.

Turban, D. B., Sanders, P. A., Francis, D. J., & Osburn, H. G. (1989). Construct equivalence as an approach to replacing validated cognitive ability selection tests. *Journal of Applied Psychology*, 74, 62-71.